Published online 29 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D901–D906***
doi:10.1093/nar/gkm958

DrugBank: a knowledgebase for drugs, drug actions and drug targets

David S. Wishart*, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam and Murtaza Hassanali

Department of Computing Science and Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E8

Received September 15, 2007; Revised October 11, 2007; Accepted October 15, 2007

ABSTRACT

DrugBank is a richly annotated resource that combines detailed drug data with comprehensive drug target and drug action information. Since its first release in 2006, DrugBank has been widely used to facilitate *in silico* drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education. The latest version of DrugBank (release 2.0) has been expanded significantly over the previous release. With ~4900 drug entries, it now contains 60% more FDA-approved small molecule and biotech drugs including 10% more 'experimental' drugs. Significantly, more protein target data has also been added to the database, with the latest version of DrugBank containing three times as many non-redundant protein or drug target sequences as before (1565 versus 524). Each DrugCard entry now contains more than 100 data fields with half of the information being devoted to drug/chemical data and the other half devoted to pharmacological, pharmacogenomic and molecular biological data. A number of new data fields, including food–drug interactions, drug–drug interactions and experimental ADME data have been added in response to numerous user requests. DrugBank has also significantly improved the power and simplicity of its structure query and text query searches. DrugBank is available at http://www.drugbank.ca

INTRODUCTION

There are essentially two kinds of online drug resources: (i) clinically oriented drug 'encyclopedias' and (ii) chemically oriented drug databases. Examples of some of the better clinically oriented drug resources include PharmGKB (1) and RxList (2). These knowledgebases tend to offer very detailed clinical information about selected drugs (their pharmacology, metabolism and indications) with their data content being targeted more towards pharmacists, physicians or consumers. Examples of chemically oriented drug (or small molecule) databases include the TTD (3), the Druggable Genome database (4), KEGG (5), PubChem (6) and ChEBI (7). These excellent databases provide synoptic data (5–10 data fields per entry) about the nomenclature, structure and/or physical properties of large numbers of small molecule drugs and, in some cases, their drug targets. Chemically oriented drug databases are typically oriented towards medicinal chemists, biochemists and molecular biologists. As a general rule, chemically oriented drug databases aim for very broad coverage at the expense of depth, while clinically oriented drug resources aim for far more depth (albeit in English sentences) at the expense of coverage.

In an effort to bridge the 'depth versus breadth' gap between clinically oriented drug resources and chemically oriented drug databases, we developed DrugBank (8). First released in 2006, DrugBank was designed to serve as a comprehensive, fully searchable *in silico* drug resource that linked sequence, structure and mechanistic data about drug molecules (including biotech drugs) with sequence, structure and mechanistic data about their drug targets. As a clinically oriented drug encyclopedia, DrugBank is able to provide detailed, up-to-date, quantitative, analytic or molecular-scale information about drugs, drug targets and the biological or physiological consequences of drug actions. As a chemically oriented drug database, DrugBank is able to provide many built-in tools for viewing, sorting, searching and extracting text, image, sequence or structure data. Since its initial release, DrugBank has been used in a wide range of applications including *in silico* drug discovery (9), drug 'rejuvenation' (10), drug docking or screening (11), drug metabolism prediction (12), drug target prediction (13) and general pharmaceutical education. Feedback from users has led to many excellent suggestions on how to expand and enhance DrugBank's offerings. These requests also led to the development of several new software tools to improve the entry, export and annotation of DrugBank's data.

*To whom correspondence should be addressed. Tel: 780-492-0383; Fax: 780-492-1071; Email: david.wishart@ulberta.ca

Here, we wish to report on these developments as well as many additions and improvements appearing in the latest version of DrugBank (release 2.0).

DATABASE ENHANCEMENTS

Details relating to DrugBank's overall design, querying capabilities, curation protocols, quality assurance and drug selection criteria have been described previously (8). These have largely remained the same between release 1.0 and 2.0. Here, we shall focus primarily on describing the changes and enhancements made to the database and to the annotation processes for release 2.0. More specifically, we will describe the: (i) enhancements to the DrugBank's size and coverage; (ii) expanded database linkages; (iii) data field additions; (iv) improvements in data querying and data viewing and (v) improvements to DrugBank's data handling processes.

Expanded database size and coverage

A detailed content comparison between DrugBank (release 1.0) versus DrugBank (release 2.0) is provided in Table 1. As seen here, the latest release of DrugBank now has detailed information on 1467 FDA-approved drugs corresponding to 28 447 brand names and synonyms. This represents an expansion of nearly 60% over what was previously contained in the database. The latest DrugBank release also includes 123 biotech (peptide or protein) drugs and 69 nutraceuticals (nutritional supplements), which corresponds to an increase of ~10% over what was in the previous DrugBank release. While many of these additions represent newly approved drugs (about 50 new drugs are approved each year), a number of these new entries are little known, hard-to-find or infrequently prescribed drugs that are not contained in most drug databases. To the best of our knowledge, DrugBank now contains all (or almost all) drugs that have been approved in North America, Europe and Asia. In addition, DrugBank's collection of experimental or unapproved drugs (or drug-like) compounds, which is primarily derived from the PDB's Ligand database, has expanded to include 3116 compounds, compared to 2896 compounds in the first release. We are pleased to note that these experimental drugs have now been more completely annotated, via BioSpider (14), than in the previous DrugBank release.

Table 1. Comparison between the data content in DrugBank (release 1.0) versus DrugBank (release 2.0)

Category	Release 1.0	Release 2.0
No. of FDA-approved small molecule drugs	841	1344
No. of biotech drugs	113	123
No. of nutraceutical drugs	61	69
No. of withdrawn drugs	0	57
No. of illicit drugs	0	188
No. of experimental drugs	2894	3116
No. of total Small molecule drugs	3796	4774
No. of total drugs	3909	4897
No. of names/brand names/synonyms	18 304	28 447
No. of data fields	88	108
No. of food–drug interactions	0	714
No. of drug–drug interactions	0	13 242
No. of ADMET parameters (Caco-2, LogS)	0	276
No. of approved drug targets (non-redundant)	524	1565
No. of all drug targets (non-redundant)	2133	3037
No. of search types	8	12

In response to many user requests, we have also added two new drug categories: (i) Withdrawn drugs and (ii) Illicit drugs. Withdrawn drugs are those that have been withdrawn from the market or certain market segments due to safety concerns (such as Vioxx and Bextra). Illicit drugs include those that are legally banned or selectively banned in most developed nations (such as cocaine and heroin). Chemical, pharmaceutical and biological information about these classes of drugs is extremely important, not only in understanding their adverse reactions, but also in being able to predict whether a new drug entity may have unexpected chemical or functional similarities to a dangerous drug. The number of drugs in the 'Withdrawn' category is 57, while the number of drugs in the 'Illicit' category is 188. As with all other entries in DrugBank, the same level of drug, drug target and drug action information has been collected for these drugs as with all other drug entries in DrugBank. If one counts all drug entries in DrugBank (FDA-approved, Experimental, Biotech, Nutraceutical, Withdrawn, Illicit), the total number of drugs or drug-like molecules comes to 4897, which represents an increase by 25% over the previous release.

A significant increase in the number (and coverage) of identified drug targets in DrugBank has been achieved for this release of DrugBank, with 1565 non-redundant protein/DNA targets being identified for FDA-approved drugs compared to 524 non-redundant targets identified in release 1.0. The identification of so many more targets was aided by PolySearch (http://wishart.biology.ualberta.ca/polysearch/), a text-mining tool developed in our laboratory to facilitate these kinds of searches. Additional details about PolySearch appear later in this article. All of these newly identified protein targets are fully referenced to an average of four PubMed citations each.

Of particular interest to many is DrugBank's list of drug targets. Several other drug target lists have been compiled or presented including those in TTD (3), as well as others by Hopkins *et al.* (15), Drews and Ryser (16), Imming *et al.* (17) and Overington *et al.* (18). These report 578 molecular targets (out of 1512 total targets including disease and organism targets), 248 protein targets (out of 399 molecular targets), 483 molecular targets, 218 molecular targets and 324 molecular targets, respectively. DrugBank's list of drug targets is 3–4 times larger than these. The primary reasons are: (i) DrugBank has a much larger collection of small molecule drugs (approximately two times larger than any other resource), (ii) DrugBank includes biotech drugs and nutraceuticals (which average 5–10 unique target proteins per drug), (iii) most other drug target lists only include a single 'primary' target rather than all targets that have been found to have physiological or pharmaceutical effects, (iv) DrugBank fully accounts for the fact that many drug targets are protein complexes

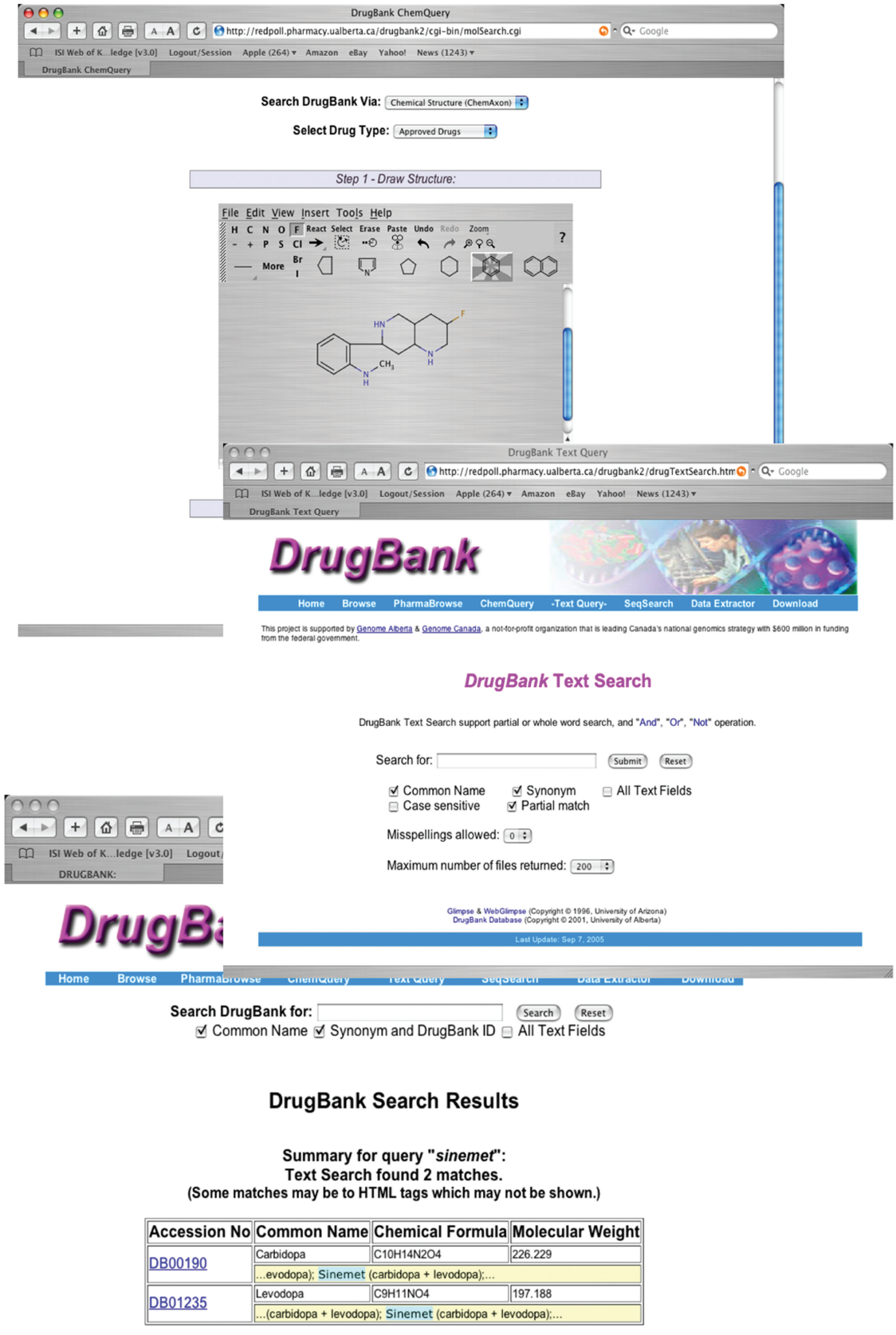

Figure 1. A screenshot montage of some of DrugBank's new or modified querying tools including ChemQuery, TextQuery and an example of the new generic text query output.

composed of multiple subunits or combinations of subunits and (v) DrugBank annotators identify molecules as drug targets if they play a critical role in the transport, delivery or activation of the drug.

As a general rule, when more than one drug target is listed in DrugBank, the ordering of the drug targets corresponds 'approximately' to their order of physiological effect or their importance regarding the drug's therapeutic indication(s).

Expanded database linkages

DrugBank is a database that contains extensive links to almost all major bioinformatics and biomedical databases (GenBank, SwissProt/UniProt, PDB, ChEBI, KEGG, PubChem and PubMed). It also contains many links to numerous drug and pharmaceutical databases (RxList, PharmGKB and FDA labels). Over the past year, DrugBank has also been reciprocally linked by SwissProt/UniProt, Wikipedia, BioMOBY (19) and PubChem (October 2007). Because of DrugBank's appeal as an educational or public information resource, we are actively seeking to expand these reciprocal linkages with other databases and online resources. For example, all drug entries in Wikipedia are now linked to DrugBank and most drug 'fact boxes' in Wikipedia are actually generated from DrugBank tables. For the latest release of DrugBank, several new database links have been added including hyperlinks to Wikipedia, PDRHealth, the Drug Product Database (DPD), the Human Genome Nomenclature Commission (HGNC), GeneCards (20) and GeneAtlas (21).

Data field additions

As seen in Table 1, DrugBank now contains 107 data fields, compared to 88 data fields in release 1.0. Some of these data fields have arisen to facilitate cataloging, but most have been added in response to user needs and user requests. Specifically, these new data fields include: (i) a primary accession number; (ii) a secondary accession number; (iii) drug synonyms; (iv) a compound description; (v) drug brand names; (vi) SwissProt name (if the drug is a peptide/protein drug); (vii) monoisotopic molecular weight; (viii) isomeric SMILES string; (ix) water solubility predicted via ALOGPS (22); (x) LogP predicted via ALOGPS; (xi) CACO permeability; (xii) experimental water solubility (LogS); (xiii) drug–drug interactions; (xiv) food–drug interactions; (xv) Human Protein Reference Database ID; (xvi) HGNC ID; (xvii) GeneCards ID and (xviii) GeneAtlas ID. A total of 194 experimental LogS values and 82 experimental Caco-2 permeability values were obtained from the UCSD ADME databases (23). These values, along with the structural and physico-chemical data in DrugBank, are particularly useful for computational ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicty) prediction. Additionally, 714 food–drug interactions and 13 242 drug–drug interactions were compiled (through a variety of web and textbook resources), checked by an accredited pharmacist and entered manually. As far as we are aware, these drug/drug and food/drug compilations represent the most complete, publicly accessible collection of its kind. This interaction information is particularly useful for physicians, pharmacists and patients. However, it is also of increasing interest to those involved in pharmacogenomics and nutrigenomics.

Enhanced querying and viewing capabilities

A key feature that distinguishes DrugBank from other online drug resources is its extensive support for higher level database searching and selecting functions. In addition to standard data viewing and sorting features, DrugBank also offers a generic text search, a local BLAST search (SeqSearch), a higher level Boolean text search (TextQuery), a chemical structure search utility (ChemQuery) and a relational data extraction tool (Data Extractor). Each of these search utilities has a number of useful bioinformatics or cheminformatic applications, many of which were described in the first DrugBank publication (8). For the latest release of DrugBank, we have added a number of improvements to both the generic text search and ChemQuery (Figure 1). In particular, the generic text search has been enhanced so that users now have the option of clicking on check boxes to limit their search to either a drug's common name, its synonyms/brand names or all text fields. Because the vast majority of queries to DrugBank are related to drug names/synonyms, the default query always has these two boxes checked off. Users wishing to search through the other 100+ data fields in DrugBank can select the 'all text fields' box. This change has also substantially improved the query response times for most DrugBank text searches.

Because the spelling of many drug names, chemical compound names and protein names is often difficult or non-intuitive, DrugBank now supports an 'intelligent' text search, where alternative spellings to misspelled or incompletely entered names are automatically provided. In addition to this change, the results from text queries have also been enhanced so that the standard tabular output (primary accession number, generic drug name, chemical formula and molecular weight) is supplemented with the query word highlighted in the selected DrugCard field(s) from which it was retrieved.

To accommodate a variety of user requests and preferences, the ChemQuery tool has been modified for release 2.0 to allow two different types of chemical drawing applets to be used: the MarvinSketch (http://www.chemaxon.com) structure drawing tool (new) and the ACD structure drawing tool (old). The MarvinSketch applet is somewhat more intuitive and easier to use, while the ChemSketch (ACD) applet is somewhat more complex but offers more structural drawing options. The default ChemQuery tool for this release is the MarvinSketch applet. DrugBank's structure querying capabilities have also been enhanced with the addition of a 'Show Similar Structure(s)' button located at the top of every DrugCard. This allows users to rapidly search for structurally similar small molecules, without having to redraw the molecule and search the database through the ChemQuery interface. Users can also limit their structure similarity search

to selected DrugBank subdatabases (Approved drugs, Nutracueticals, Illicit drugs, etc.) through a pull-down menu located by the 'Show Similar Structure(s)' button. Both 'Show Similar Structures' and ChemQuery use a locally developed SMILES string comparison method to identify related structures and to perform structure similarity searches. All structures are converted to SMILES strings and a substring-matching program (similar to BLAST) is used to identify similar structures. The scoring scheme is based simply on the number of character matches for the longest matching substring.

Improved data handling (entry, export and annotation)

For most of the past 5 years, DrugBank has existed as a series of text files that were manually edited or flat files that were populated by writing Perl scripts to reformat existing text to the DrugBank file format. Most of the annotation in DrugBank (release 1.0) was assembled, entered and validated manually. With the rapid growth in the size and scope of DrugBank, along with the continuing needs for updates, we have had to become far more efficient in our data management. Specifically, we have had to streamline our methods for data entry, data export and database annotation. However, we have continued to maintain our same rigorous standards for manual data validation.

To facilitate manual data entry and export for release 2.0, we have developed customized scientific data management software (SDMS) called DrugBank–SDMS. This web-enabled database system was built using the open source Ruby-on-Rails web application framework. This SDMS overlays a MySQL database that contains all of the DrugBank data. The publicly viewable version of DrugBank is directly linked to the DrugBank–SDMS such that every night the SDMS data is automatically exported to the DrugBank server. This 'near synchrony' between the SDMS and DrugBank allows our database annotators to remotely access the SDMS, to add data, to check entries or to make corrections in real time, without the need to write (or wait for) custom Perl scripts for data uploads. The use of a SDMS also allows for more extensive error checking. This is done both at the time of entry (via automated format and spelling checks) and later (once a week), through the use of 'sanity checker' (Supplementary Table 1) that checks the consistency of chemical structure files, chemical formulae and chemical properties using a variety of custom-built prediction and file-formatting programs (8, 14, 24). The development of a custom SDMS has also facilitated the export of publicly downloadable DrugBank files. In particular, our SDMS allows rapid generation of all of DrugBank's flat file (text) downloads and facile creation of XML-formatted DrugBank files—all of which are available through DrugBank's download link.

To improve our manual annotation efficiency and coverage, the programming staff at DrugBank has developed several automated text and web-mining tools including BioSpider (14) and PolySearch. BioSpider is a web spider that automatically gathers biological, chemical and pharmacological data from approximately 30 trusted, content-rich web sites using only a compound name, SMILES string or Chemical Abstract Service (CAS) number as input. It then combines this data with a variety of in-house molecular structure and property prediction tools to generate data tables that corresponds to many of the data fields in DrugBank. BioSpider allows many of the tedious, error-prone or repetitive annotation activities in DrugBank to be handled by a computer, allowing our annotation team to concentrate on higher level annotation tasks (such as, gathering data on pharmacology, mechanism of action, metabolism or drug interactions). BioSpider has been extensively evaluated (14) and has been found to perform much better and much faster than skilled human annotators in these low-level annotation tasks. To complement BioSpider's role in low-level annotation, we have also developed PolySearch to enhance higher level annotation and research. PolySearch is a text-mining tool designed to mine data from abstracts in PubMed. It is similar in concept and design to EBIMed (25) and MedGene (26), but has been modified to facilitate the extraction of informative sentences or informative abstracts related to drugs, drug targets, drug metabolites, diseases, proteins and drug–protein interactions. PolySearch is used as an adjunct to our manual annotation efforts and has greatly aided the identification of numerous or little-known drug targets.

All textual data acquired from the BioSpider and PolySearch annotation programs are manually inspected by a minimum of two individuals, with at least one individual having an MD or a life science PhD. Additional spot checks are routinely performed on each entry by senior members of the curation group, including a physician, an accredited pharmacist and two PhD-level biochemists. While most information listed in the 'Drug Description', 'Pharmacology', 'Mechanisms of Action', 'Half Life', 'Biotransformation Data', 'Protein Binding', 'Toxicity', 'Absorption' and 'Indications' data fields is manually entered, those entries that are acquired from our automated annotation tools are all manually verified and edited (or rewritten) for readability and consistency. All PolySearch-derived drug target data, in particular, has been verified through multiple text sources (PubMed, drug references, online sequence databases, online drug databases and FDA labels) by at least two members of the DrugBank curatorial staff. Drugs with near-identical structures and modes of action are cross-checked to ensure that their drug target lists are nearly identical. In addition to these manual checks, nearly 40 automated data consistency checks are performed to ensure a uniformly high level of data integrity (Supplementary Table 1). Even with these added checks and references we still recommend that users carefully study the data sources prior to making decisions about using it.

FUTURE DIRECTIONS

The DrugBank model of 'breadth + depth' has served as a good template for the development of other small molecule databases in our laboratory, including the Human Metabolome Database or HMDB (24) and

FooDB (http://hmdb.med.ualberta.ca/foodb). The lessons learned from building these and other related 'metabolomic' databases are also helping to generate ideas, software and protocols that could significantly enhance the breadth and depth of information contained in future releases of DrugBank. Over the coming 3 years, DrugBank will adhere to a semi-annual updating schedule with new updates being released on the January 1 and July 1 of each year. This will allow information on newly approved and newly withdrawn drugs to be kept current. Previous versions of the database will be available from the DrugBank download page. A major focus over the coming 2 years will be to extend the database's querying capabilities (improved structure searches), to acquire more experimental spectral (MS and NMR) data, to expand its coverage of nutraceuticals or herbal medicines, to enhance the annotation of research/experimental compounds, to add many more pathway or network diagrams and to add a number of Java plug-ins to facilitate virtual drug screening and pharmacological (ADMET) modeling.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank the Canadian Institutes for Health Research (CIHR), as well as Genome Alberta and Genome Canada for financial support. We are also indebted to the many users of DrugBank who have provided valuable feedback and suggestions. Funding to pay the Open Access publication charges was provided by Genome Alberta.

Conflict of interest statement: None declared.

REFERENCES

1. Hodge,A.E., Altman,R.B. and Klein,T.E. (2007) The PharmGKB: integration, aggregation, and annotation of pharmacogenomic data and knowledge. *Clin. Pharmacol. Ther.*, **81**, 21–24.
2. Hatfield,C.L., May,S.K. and Markoff,J.S. (1999) Quality of consumer drug information provided by four web sites. *Am. J. Health Syst. Pharm.*, **56**, 2308–2311.
3. Chen,X., Ji,Z.L. and Chen,Y.Z. (2002) TTD: therapeutic target database. *Nucleic Acids Res.*, **30**, 412–415.
4. Russ,A.P. and Lampel,S. (2005) The druggable genome: an update. *Drug Discov. Today*, **10**, 1607–1610.
5. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34** (Database issue), D354–D357.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35** (Database issue), D5–D12.
7. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33** (Database issue), D46–D53.
8. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34** (Database issue), D668–D672.
9. Chang,C., Bahadduri,P.M., Polli,J.E., Swaan,P.W. and Ekins,S. (2006) Rapid identification of P-glycoprotein substrates and inhibitors. *Drug Metab. Dispos.*, **34**, 1976–1984.
10. Chong,C.R., Sullivan, and D.J., Jr (2007) New uses for old drugs. *Nature*, **448**, 645–646.
11. Li,H., Gao,Z., Kang,L., Zhang,H., Yang,K., Yu,K., Luo,X., Zhu,W., Chen,K. *et al.* (2006) TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.*, **34** (Web Server issue), W219–W224.
12. Jolivette,L.J. and Ekins,S. (2007) Methods for predicting human drug metabolism. *Adv. Clin. Chem.*, **43**, 131–176.
13. Wishart,D.S. (2007) Discovering drug targets through the web. *Comp. Biochem. Physiol. D*, **2**, 9–17.
14. Knox,C., Shrivastava,S., Stothard,P., Eisner,R. and Wishart,D.S. (2007) BioSpider: a web server for automating metabolome annotations. *Pac. Symp. Biocomput.* 145–156.
15. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
16. Drews,J. and Ryser,S. (1997) The role of innovation in drug development. *Nat. Biotechnol.*, **15**, 1318–1319.
17. Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.*, **5**, 821–834.
18. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L. (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
19. Kawas,E., Senger,M. and Wilkinson,M.D. (2006) BioMoby extensions to the Taverna workflow management and enactment software. *BMC Bioinformatics*, **7**, 523.
20. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.
21. Kitson,D.H., Badretdinov,A., Zhu,Z.Y., Velikanov,M., Edwards,D.J., Olszewski,K., Szalma,S. and Yan,L. (2002) Functional annotation of proteomic sequences based on consensus of sequence and structural analysis. *Brief. Bioinform.*, **3**, 32–44.
22. Tetko,I.V and Tanchuk,V.Y. (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.*, **42**, 1136–1145.
23. Hou,T., Wang,J., Zhang,W., Wang,W. and Xu,X. (2006) Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.*, **13**, 2653–2667.
24. Wishart,D.S., Tzur,D., Knox,C., Eisner,R., Guo,A.C., Young,N., Cheng,D., Jewell,K., Arndt,D. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.*, **35** (Database issue), D521–D526.
25. Rebholz-Schuhmann,D., Kirsch,H., Arregui,M., Gaudan,S., Rynbeek,M. and Stoehr,P. (2006) Protein annotation by EBIMed. *Nat. Biotechnol.*, **24**, 902–903.
26. Hu,Y., Hines,L.M., Weng,H., Zuo,D., Rivera,M., Richardson,A. and LaBaer,J. (2003) Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res.*, **2**, 405–412.